Understanding Neural Machine Translation by Simplification: The Case of Encoder-free Models

Understanding Neural Machine Translation by Simplification: The Case of Encoder-free Models

Gongbo Tang[1] **Rico Sennrich**[2,3] **Joakim Nivre**[1]

[1]Department of Linguistics and Philology, Uppsala University
[2]School of Informatics, University of Edinburgh
[3]Institute of Computational Linguistics, University of Zurich
firstname.lastname@{lingfil.uu.se, ed.ac.uk}

Abstract

In this paper, we try to understand neural machine translation (NMT) via simplifying NMT architectures and training encoder-free NMT models. In an encoder-free model, the sums of word embeddings and positional embeddings represent the source. The decoder is a standard Transformer or recurrent neural network that directly attends to embeddings via attention mechanisms. Experimental results show (1) that the attention mechanism in encoder-free models acts as a strong feature extractor, (2) that the word embeddings in encoder-free models are competitive to those in conventional models, (3) that non-contextualized source representations lead to a big performance drop, and (4) that encoder-free models have different effects on alignment quality for German→English and Chinese→English.

1 Introduction

Neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015) has emerged in the last few years and has achieved new state-of-the-art performance. However, NMT models are black boxes for humans and are hard to interpret. NMT models employ encoder-decoder architectures where an encoder encodes source-side sentences and an attentional decoder generates target-side sentences based on the outputs of the encoder. In this paper, we attempt to obtain a more interpretable NMT model by simplifying the encoder-decoder architecture. We train encoder-free models where the sums of word embeddings and sinusoid embeddings (Vaswani et al., 2017) represent the source. The decoder is a standard Transformer (Vaswani et al., 2017) or recurrent neural network (RNN) that attends to embeddings via attention mechanisms.

As motivation for our architecture simplification, consider the attention mechanism[1] (Bahdanau et al., 2015; Luong et al., 2015), which has been introduced to extract features from the hidden representations in encoders dynamically. Attention and alignment were initially used interchangeably, but it was soon discovered that the attention mechanism can behave very differently from traditional word alignment (see Ghader and Monz, 2017; Koehn and Knowles, 2017). One reason for this discrepancy is that the attention mechanism operates on representations that potentially includes information from the whole sentence due to the encoder's recurrent or self-attentional architecture. Intuitively, bypassing these encoder layers and attending word embeddings directly could lead to a more alignment-like, and thus predictable and interpretable behavior of the attention model.

By comparing encoder-free models with conventional models, we can better understand the working mechanism of NMT, figure out which components are more crucial, and learn lessons for improvement. Experimental results show that there is a significant gap between the two models. We focus on exploring what leads to the big gap.

As the embeddings in encoder-free Transformers (*Trans-noEnc*) are only influenced by attention mechanisms, without the help of encoders, we hypothesize that the quality of embeddings leads to the gap between Transformers and *Trans-noEnc* models. Thus we conduct both qualitative and quantitative evaluations of the embeddings from Transformers and *Trans-noEnc* models. We also hypothesize that the attention distribution in *Trans-noEnc* is not spread out enough for extract-

[1]We refer to the encoder-decoder attention mechanism unless otherwise specified.

ing contextual features. However, we find that word embeddings and attention distributions are not the major reasons causing the distinct gap. We further explore NMT encoders. We find that even NMT models with one layer encoder get significant improvement compared to encoder-free models which indicates that non-contextualized source representations lead to the evident gap.

In encoder-free models, the attention attends to source embeddings rather than hidden representations fused with the context. We hypothesize that encoder-free models generate better alignments than default models. We evaluate the alignments generated on German→English (DE→EN) and Chinese→English (ZH→EN). We find that encoder-free models improve the alignments for DE→EN but worsen the alignments for ZH→EN.

2 Related Work

2.1 Understanding NMT

The attention mechanism has been introduced as a way to learn an alignment between the source and target text, and improves encoder-decoder models significantly, while also providing a way to interpret the inner workings of NMT models. However, Ghader and Monz (2017) and Koehn and Knowles (2017) have shown that the attention mechanism is different from a word alignment. While there are linguistically plausible explanations in some cases – when translating a verb, knowledge about the subject, object etc. may be relevant information – other cases are harder to explain, such as an off-by-one mismatch between attention and word alignment for some models. We suspect that such a pattern can be learned if relevant information is passed to neighboring representations via recurrent or self-attentional connections.

Ding et al. (2017) show that only using attention is not sufficient for deep interpretation and propose to use layer-wise relevance propagation to better understand NMT. Wang et al. (2018) replace the attention model with an alignment model and a lexical model to make NMT models more interpretable. The proposed model is not superior but on a par with the attentional model. They clarify the difference between alignment models and attention models by saying that that the alignment model is to identify translation equivalents while the attention model is to predict the next target word.

In this paper, we try to understand NMT by simplifying the model. We explore the importance of different NMT components and what causes the performance gap after model simplification.

2.2 Alignments and Source Embeddings

Nguyen and Chiang (2018) introduce a lexical model to generate a target word directly based on the source words. With the lexical model, NMT models generate better alignments. Kuang et al. (2018) propose three different methods to bridge source and target word embeddings. The bridging methods can significantly improve the translation quality. Moreover, the word alignments generated by the model are improved as well.

Our encoder-free model is a simplification and only attends to the source word embeddings. We aim to interpret NMT models rather than pursuing better performance.

Different from previous work, Zenkel et al. (2019) introduce a separate alignment layer directly optimizing the word alignment. The alignment layer is an attention network learning to attend to source tokens given a target token. The attention network can attend to either the word embeddings or the hidden representations or both of them. The proposed model significantly improves the alignment quality and performs as well as the aligners based on traditional IBM models.

3 Experiments

In addition to training Transformer and *Trans-noEnc* models, we also compare *Trans-noEnc* with NMT models based on RNNs (*RNNS2S*). We train *RNNS2S* models without encoders (*RNNS2S-noEnc*), without attention mechanisms (*RNNS2S-noAtt*), and without both encoders and attention mechanisms (*RNNS2S-noAtt-noEnc*) to explore which component is more important for NMT. We also investigate the importance of positional embeddings in *Trans-noEnc*.

3.1 Experimental Settings

We use the *Sockeye* (Hieber et al., 2017) toolkit, which is based on MXNet (Chen et al., 2015), to train models. Each encoder/decoder has 6 layers. For *RNNS2S*, we choose long short-term memory (LSTM) RNN units. Transformers have 8 attention heads. The size of embeddings and hidden states is 768. We tie the source, target, and output embeddings. The dropout rate of embeddings and Transformer blocks is set to 0.1. The dropout rate

of RNNs is 0.2. All the models are trained with a single GPU. During training, each mini-batch contains 2,048 tokens. A model checkpoint is saved every 1,000 updates. We use *Adam* (Kingma and Ba, 2015) as the optimizer. The initial learning rate is set to 0.0001. If the performance on the validation set has not improved for 8 checkpoints, the learning rate is multiplied by 0.7. We set the early stopping patience to 32 checkpoints.

The training data is from the WMT15 shared task (Bojar et al., 2015) on Finnish–English (FI–EN). We choose *newsdev2015* as the validation set and use *newstest2015* as the test set. All the BLEU (Papineni et al., 2002) scores are measured by *SacreBLEU* (Post, 2018). There are about 2.1M sentence pairs in the training set after preprocessing. We learn a joint BPE model with 32K subword units (Sennrich et al., 2016). We employ the models that have the best perplexity on the validation set for the evaluation. We set the beam size to 8 during inference.

To test the universality of our findings, we conduct experiments on DE→EN and ZH→EN as well. For DE→EN, we use the training data from the WMT17 shared task (Bojar et al., 2017). We use *newstest2013* as the validation set and *newstest2017* as the test set. We learn a joint BPE model with 32k subword units. For ZH→EN, we choose the CWMT parallel data of the WMT17 shared task for training. We use *newsdev2017* as the validation set and *newstest2017* as the test set. We apply Jieba[2] to Chinese segmentation. We then learn 60K subword units for Chinese and English separately. There are about 5.9M and 9M sentence pairs in the training set after preprocessing in DE→EN and ZH→EN, respectively.

3.2 Results

Table 1 shows the performance of all the trained models. Encoder-free models (*NMT-noEnc*s) perform rather poorly compared to conventional NMT models.[3] It is interesting that *Trans-noEnc* obtains a BLEU score similar to the *RNNS2S* model. Even though the attention networks only attend to the non-contextualized word embeddings, *Trans-noEnc* still performs as well as the *RNNS2S* by paying attention to the context with multiple attention layers. Tang et al. (2018a) find that the superiority of Transformer models is attributed to the self-attention network which is a powerful semantic feature extractor. Given our results, we conclude that the attention mechanism is also a strong feature extractor in *Trans-noEnc* without self-attention in the encoder.

Model	Param.	PPL	BLEU
Transformer	104.4M	9.6	18.9
Trans-noEnc	71.4M	11.7	15.9
RNNS2S	91.5M	14.9	15.9
RNNS2S-noEnc	64.3M	25.2	12.5
RNNS2S-noAtt	90.3M	33.3	8.2
RNNS2S-noAtt-noEnc	63.1M	53.7	4.1
Trans-noEnc-noPos	71.4M	26.6	7.1

Table 1: The performance of NMT models. PPL is the perplexity on the development set. BLEU scores are evaluated on *newstest2015*. "Param." denotes the number of parameters.

The attention mechanism improves encoder-decoder architectures significantly. However, there are no empirical results to clarify whether encoders or attention mechanisms are more important for NMT models. We compare *RNNS2S-noAtt*, *RNNS2S-noEnc*, and *RNNS2S-noAtt-noEnc* to explore which component contributes more to NMT models.[4] In Table 1, *RNNS2S-noEnc* performs much better than *RNNS2S-noAtt*. Moreover, the gap between *RNNS2S-noEnc* and *RNNS2S-noAtt-noEnc* is distinctly larger than the gap between *RNNS2S-noAtt* and *RNNS2S-noAtt-noEnc*. These results hint that attention mechanisms are more powerful than encoders in NMT.

The positional embedding is also very important to Transformers which holds the sequential information. We are interested in the extent to which the positional embedding affects the translation performance. We further simplify the model by removing the positional embedding in the source (*Trans-noEnc-noPos*). *Trans-noEnc-noPos* has a dramatic drop in BLEU score. It is even worse than *RNNS2S-noAtt*. This result indicates that positional information is indeed crucial for Transformers.

[2] https://github.com/fxsjy/jieba

[3] We also trained a *Transformer* with less parameters (64.3M). The *Transformer* still achieved a significantly better BLEU score (18.2) than *Trans-noEnc* which means that the number of parameters is not the primary factor in this case.

[4] Because the encoders and decoders in Transformers are only connected via attention, we only conduct this experiment on *RNNS2S* models.

Word	Neighbors	
	Transformer	*Trans-noEnc*
more	less, better, greater, most, **further**	less, greater, better, **fewer**, most
for	to, in, on, of, **with**	to, in, of, on, **towards**
ole (not)	olekaan (not the), **kykene** (unable to), kuulu (part of), **pysty** (upright), **ollut** (been)	olekaan, kuulu (part of), **ei** (no/not), **ene** (a suffix), **liity** (sign up)
Arvoisa (honorable)	arvoisa, Arvoisat (honorable), **arvoisaa**, **arvoisan** (honorable), hyvät (honorable)	arvoisa, **arvoisat**, hyvät, Arvoisat, **Hyvä** (honorable)

Table 2: Neighbors of the selected word embeddings. Bold words are distinct neighbors.

4 Analysis

Trans-noEnc is obviously inferior to *Transformer* but we are more interested in investigating what causes the performance gap. In this section, we will test our hypotheses on embedding quality and attention distributions.

4.1 Embeddings

Word embeddings are randomly initialized by default and learned during training. As the embeddings in *Trans-noEnc* are only updated by attention mechanisms, we hypothesize that embeddings in *Trans-noEnc* are not well learned and therefore affect translation performance. We test our hypothesis by (1) evaluating the embeddings in the two models manually and (2) initializing *Trans-noEnc* with the learned embeddings in *Transformer* as pre-trained embeddings.

Qualitative Evaluation We select the 150 most frequent tokens from the vocabulary and then manually evaluate the quality of embeddings by comparing the 5 nearest neighbors.

The quality of English word embeddings is quite good based on the output of neighbors. Finnish word embeddings are not as good as English word embeddings. Table 2 exhibits four examples, two English words, "more", "for" and two Finnish word, "ole" (not), "Arvoisa" (honorable). The neighbors of "more" in *Transformer* and *Trans-noEnc* are all quite related words, including comparatives and "most" which is the superlative of "more". The words "further" and "fewer" are more different neighbors but both are related to "more". For the Finnish word "ole" (not), both models have negative words as neighbors, but there are different unrelated words as well. We can see that the qualities of neighbors in two embedding matrices are close. We cannot easily distinguish which embedding matrix is better based on the neighbors.

Quantitative Evaluation In addition to the qualitative evaluation, we also conduct a quantitative evaluation. We first employ the learned embeddings from *Transformer* to initialize the embedding parameters in *Trans-noEnc*. The pre-trained embeddings can be either fixed or not fixed during training. Table 3 gives the BLEU scores of these models. The pre-trained embeddings slightly improve the BLEU score.

Embeddings	Random	Fixed	Not-fixed
BLEU	15.9	16.1	16.2

Table 3: BLEU scores of *Trans-noEnc*s with different embedding initialization. "Random" means no pre-trained embeddings. "Fixed" and "Not-fixed" denote using pre-trained embeddings.

The evaluation reveals that the embeddings from *Trans-noEnc* are competitive to those of *Transformer*. Thus, we can rule out differences in embedding quality as the main factor for the performance drop.

4.2 Attention Distribution

The attention networks in *Trans-noEnc* only attend to word embeddings. To better capture the sentence-level context, the attention networks need to distribute more attention to the context. We test our hypothesis that the attention distributions in *Trans-noEnc* are not as distributed as those in *Transformer*. If the attention distributions in *Transformer* are more spread out than those in *Trans-noEnc*, it means that smaller weights are distributed to contextual features by *Trans-noEnc*.

$$E_{At}(y_t) = -\sum_{i=1}^{|x|} At(x_i, y_t) \log At(x_i, y_t) \quad (1)$$

We use attention entropy (Equation 1) (Ghader and Monz, 2017) to measure the concentration of the attention distribution at timestep t. We then average the attention entropy at all the timesteps as the final attention entropy. x_i denotes the ith source token, y_t is the prediction at timestep t, and $At(x_i, y_t)$ represents the attention distribution at timestep t. The attention mechanism in Transformer has multiple layers, and each layer has multiple heads. In each layer, we average the attention weights from all the heads.

Figure 1 shows the entropy of attention distributions in both models. The attention distributions are consistent with the finding in Tang et al. (2018b) that the distribution gets concentrated first and then becomes distributed again. *Transformer* has lower entropy, which potentially is because the contextual information has been encoded in the hidden representations. The attention entropy of *Trans-noEnc* is clearly higher than that of *Transformer* in each attention layer. The attention in *Trans-noEnc* tends to extract features from source tokens more uniformly which indicates that the attention mechanism compensates for the fact that embeddings are non-contextualized by distributing attention across more tokens.

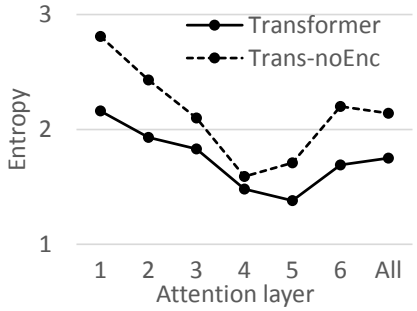

Figure 1: The attention entropy of each attention layer and the entire attention mechanism.

4.3 Encoders

We have shown that embeddings and attention distributions are not the primary reasons causing the gap between *Transformer* and *Trans-noEnc*. Therefore, we move to explore encoders.

Encoders are responsible for providing source hidden representations to the decoder. Encoder-free models have to use word embeddings to represent source tokens without the help of encoders. Thus, the source-side representations probably lead to the performance gap.

We train NMT models with different encoder layers. Table 4 displays the performance of Transformer models that have different layers in the encoder. It is clear that even the model with only a 1-layer encoder outperforms *Trans-noEnc* (0-layer) by 1.7 BLEU points, which accounts for 56.7% of the performance gap. The results seem to show that source-side hidden representations are crucial in NMT.

Layers	Param.	PPL	BLEU
0	71.4M	11.7	15.9
1	76.9M	10.3	17.6
3	87.9M	9.9	18.4
5	98.9M	9.5	18.6
6	104.4M	9.6	18.9

Table 4: The performance of Transformer models that have different layers in the encoder, including the perplexity (PPL) on the development set and the BLEU scores on *newstest2015*.

It has been shown that encoders could extract syntactic and semantic features in NMT (Belinkov et al., 2017a,b; Poliak et al., 2018). In the meantime, contextual information is encoded in hidden representations as well. Hence we conclude that the quality of source representations is the main factor causing the big gap between *Transformer* and *Trans-noEnc*.

In Table 5, our additional experiments on DE→EN and ZH→EN confirm that models with contextualized representations are much better. Transformer models always outperform *Trans-noEnc* models substantially.

Lan.	*Trans-noEnc*	*Transformer*	Impr.
DE→EN	29.5	32.6	10.5%
ZH→EN	18.5	20.9	13.0%

Table 5: The improvement (Impr.) of employing encoders in *Trans-noEnc*s on DE→EN and ZH→EN.

5 Alignment

The weights of the attention mechanism can be interpreted as an alignment between the source and target text. We further explore whether encoder-free models have better alignments than default models. We evaluate the alignments on two manually annotated alignment data sets. The first one

has been provided by RWTH,[5] and consists of 508 DE→EN sentence pairs. The other one is from Liu and Sun (2015) and contains 900 ZH→EN sentence pairs. We apply alignment error rate (AER) (Och and Ney, 2003) as the evaluation metric.

Following Luong et al. (2015); Kuang et al. (2018), we also force the models to produce the reference target words during inference to get the alignment between input sentences and their reference outputs. We merge the subwords after translation following the method in Koehn and Knowles (2017).[6] We sum the attention weights in all attention heads in each attention layer.[7] Given a target token, the source token with the highest attention weight is viewed as the alignment of the current target token (Luong et al., 2015). However, a source token maybe aligned to multiple target tokens and vice versa. Therefore, we also align a source token to the target token that has the highest attention weight given the source token. Experimental results show that the bidirectional method achieves higher alignment quality.

Figure 2 displays the evaluation results. The alignment in the fourth attention layer achieves the best performance. Therefore, we only compare the alignments in the fourth layer. In DE→EN, the encoder-free model has a lower AER score (0.41) than the default model (0.43) which accords with our hypothesis. However, in ZH→EN, the alignment quality of the encoder-free model (0.46) is worse than that of the default model (0.43). The effect on alignment quality is not clear-cut for encoder-free models given limited language pairs.

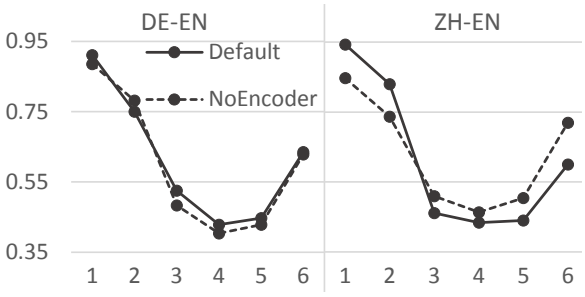

Figure 2: The AER scores of alignments in different attention layers on DE→EN and ZH→EN.

[5] https://www-i6.informatik.rwth-aachen.de/goldAlignment/

[6] (1) If an input word is split into subwords, we sum their attention weights. (2) If a target word is split into subwords, we average their attention weights.

[7] Following Tang et al. (2018b), we tried maximizing the attention weights as well but got worse alignment quality.

6 Conclusion

To better understand NMT, we simplify the attentional encoder-decoder architecture by training encoder-free NMT models in this paper. The non-contextualized source representations in encoder-free models cause a big performance drop, but the word embeddings in encoder-free models are shown competitive to those in default models. Also, we find that the attention component in encoder-free models is a powerful feature extractor, and can partially compensate for the lack of contextualized encoder representations.

Regarding the interpretability of attention, our results do not show that the attention mechanism in encoder-free models is consistently more alignment-like: only attending to source embeddings improves the alignment quality on DE→EN but makes the alignment quality worse on ZH→EN.

Acknowledgments

We thank all reviewers for their valuable and insightful comments. Gongbo Tang is mainly funded by the Chinese Scholarship Council (grant number 201607110016).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, California, USA. https://arxiv.org/abs/1409.0473.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 861–872. http://aclweb.org/anthology/P17-1080.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 1–10. http://www.aclweb.org/anthology/I17-1001.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang,

Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, pages 169–214. http://aclweb.org/anthology/W17-4717.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 1–46. http://aclweb.org/anthology/W15-3001.

Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. In *Proceedings of the Workshop on Machine Learning Systems in Neural Information Processing Systems 2015*. http://arxiv.org/abs/1512.01274.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1150–1159. https://doi.org/10.18653/v1/P17-1106.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pages 30–39. http://www.aclweb.org/anthology/I17-1004.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690* http://arxiv.org/abs/1712.05690.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1700–1709. http://www.aclweb.org/anthology/D13-1176.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, California, USA. https://arxiv.org/abs/1412.6980.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, Vancouver, Canada, pages 28–39. http://www.aclweb.org/anthology/W17-3204.

Shaohui Kuang, Junhui Li, António Branco, Weihua Luo, and Deyi Xiong. 2018. Attention focusing for neural machine translation by bridging source and target embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pages 1767–1776. http://aclweb.org/anthology/P18-1164.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Austin, Texas, USA, pages 2295–2301.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1412–1421. http://aclweb.org/anthology/D15-1166.

Toan Nguyen and David Chiang. 2018. Improving lexical choice in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, USA, pages 334–343. http://aclweb.org/anthology/N18-1031.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51. http://www.aclweb.org/anthology/J03-1002.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://www.aclweb.org/anthology/P02-1040.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018. On the evaluation of semantic phenomena in neural machine translation using natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*

(Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, USA, pages 513–523. https://doi.org/10.18653/v1/N18-2082.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 186–191. http://aclweb.org/anthology/W18-6319.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Neural Information Processing Systems 2014*. Montréal, Canada, pages 3104–3112. https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018a. Why self-attention? a targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 4263–4272. http://aclweb.org/anthology/D18-1458.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018b. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 26–35. http://aclweb.org/anthology/W18-6304.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pages 6000–6010. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Weiyue Wang, Derui Zhu, Tamer Alkhouli, Zixuan Gan, and Hermann Ney. 2018. Neural hidden markov model for machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, pages 377–382. https://www.aclweb.org/anthology/P18-2060.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359* https://arxiv.org/abs/1901.11359.